STATISTICAL METHODOLOGY - ATS

INTRODUCTION

The Bureau of the Census conducts the Annual Trade Survey (ATS) to provide national estimates of annual sales, e-commerce sales, and other measures of economic activity of wholesale establishments by kind of business.

We develop the estimates in this report using data from a probability sample and administrative records. The sample is taken from a universe of firms which are engaged in wholesale trade in the United States and have paid employees. Firms of all sizes and wholesale kinds of business are selected.

This is the first time that e-commerce sales data were collected in the ATS. The data was collected by adding questions to all ATS questionnaires and mailed to all firms in the sample. Similar methods were used to estimate total sales and e-commerce sales.

STATISTICAL METHODOLOGY

This section describes the statistical methodology used in the Annual Trade Survey.

Sampling Frame

A new sample was used to compute estimates based on the North American Industry Classification System (NAICS) beginning with the 1998 reference year. This section describes the design and selection procedures for the new sample.

The sampling frame for the Annual Trade Survey (ATS) has two types of sampling units represented -- Employer Identification Numbers and large, multiple-establishment firms. Both sampling units represent clusters of one or more establishments owned or controlled by the same firm. All of the information used to create these sampling units was extracted from establishment records contained on the Census Bureau's Business Register as updated to June 1999. The next few paragraphs give details about the Business Register; the distinction between firms, EINs, and establishments; and the construction of the sampling units used in the sample selection.

The Business Register is a multi-relational database that contains a record for each establishment with paid employees. The establishment is the smallest entity represented on the Business Register. An *establishment* is a single physical location where business transactions take place and for which payroll and employment records are kept. Groups of one or more establishments under common ownership or control are *firms*. A *singleunit* firm owns or operates only one establishment, whereas a *multiunit* firm owns or operates two or more establishments. The treatment of establishments on the Business Register differs according to whether the establishment is part of a multiunit or singleunit firm. In particular, the structure of an establishment’s primary identifier on the Business Register differs depending on whether it is owned by a *singleunit* firm or by a *multiunit* firm.

A singleunit firm’s primary identifier is its Employer Identification Number (EIN). The Internal Revenue Service issues the EIN and the firm uses it as an identifier to report social security payments for its employees under the Federal Insurance Contributions Act (FICA). The same act requires all employer firms to use EINs. Because singleunit firms have only one establishment, there is a one-to-one relationship between the firm and the EIN. Thus the firm, the EIN, and the establishment all reference the same physical location and all three terms can be used interchangeably and unambiguously when referring to single establishment firms.

For multiunit firms however, a different structure connects an employer firm with its establishments via the EIN. Essentially a multiunit firm is associated with a cluster of one or more EINs and EINs are associated with two or more establishments. Each firm is associated with at least one EIN and only one firm can use a given EIN. However, one firm may have several EINs. Similarly, there is a one-to-many relationship between EINs and establishments. Each EIN can be associated with many establishments but each establishment is associated with only one EIN. Because of the possibility of one-to-many relationships, we must distinguish between the firm, its EINs, and its establishments. The firm that owns or controls a multiunit establishment is identified on the Business Register by way of the establishment’s primary identifier.

A multiunit establishment’s primary identifier consists of a unique combination of an alpha number and a plant number. The alpha number identifies the firm and the plant number identifies the establishment within the firm. All establishments owned or controlled by the same firm have the same alpha number. Different firms have different alpha numbers and different establishments within the same firm have different plant numbers. The Census Bureau assigns both the alpha number to the firm and the plant numbers to the establishments based on the results of the quinquennial economic censuses and the annual Company Organization Surveys.

To create the sampling frame, we extract the records for all establishments that are classified according to the North American Industry Classification System (NAICS) in the Wholesale Trade sector. The NAICS is defined in the North American Industry Classification, United States, 1997. For these establishments, we extracted sales (or receipts), payroll, employment, inventory, name and address information, as well as primary identifiers and, for multiunit establishments, associated EINs. We summarize the economic data of multiunit establishments to an EIN level by tabulating the establishment data for all wholesale establishments associated with the same EIN. Similarly we summarize to a firm level by tabulating the establishment data for all establishments associated with the same alpha number. These are the sampling units created from multiunit establishments. No aggregation is necessary to put singleunit establishment information on an EIN basis or a firm basis. Thus, the sampling units created for singleunit firms simultaneously represent establishment, EIN, and firm information.

Stratification, Sampling Rates, and Allocation

The primary stratification of the frame is by kind-of-business group based on the detail required for the ATS. We further stratify the sampling units within kind of business groups (substratify) by a measure of size related to their sales. To reduce the variance of the estimates, the largest sampling units are selected “with certainty”. This means they are sure to be selected and will represent only themselves (i.e., have a selection probability of one and a sampling weight of one). Within each kind of business stratum a substratum boundary (or cutoff) that divide the certainty units from the noncertainty units is determined. We based these cutoffs on a statistical analysis of the 1997 Census of Wholesale Trade. Accordingly, these values are on a 1997 sales basis. We also used this analysis to set sampling rates needed to achieve specified sampling variability objectives for sales and inventory estimates for different kind-of-business groups. Note that we computed sampling rates using data from the 1997 Census, the actual sample size was determined by applying these sampling rates to the sampling frame constructed from the 1999 Business Register. We then allocated the sample optimally based on the number of sampling units and standard deviation of the units’ measures of size. The allocation is optimal in the sense that it achieves the minimum variance for a fixed sample size.

Sample Selection

The first step in the sample selection identified certainty firms. If a firm had total merchant wholesale sales or inventories (for 1999 adjusted to a 1997 basis) greater than the corresponding certainty cutoff for its major kind of business, the firm was selected into the sample with certainty. In this case, the firm is considered the sampling unit. This has important consequences when the firm adds establishments. If a firm was selected with certainty and had more than one establishment at the time of sampling, any new establishments that the firm acquires, even if under new or different EINs, are included in the sample with certainty. If a singleunit firm was selected with certainty only future establishments associated with that firm are included with certainty; any new EINs that might later be associated with that firm are subjected to sampling as births (see below).

All firms not selected with certainty were subjected to sampling on an EIN basis. If a firm had more than one EIN, we treated each of its EINs as a separate sampling unit. To be eligible for the initial sampling, an EIN had to have nonzero payroll in 1998. EINs used by multiunits were required to have nonzero payroll in 1997. The EINs were stratified according to their major kind of business and their estimated sales (on a 1997 basis). Within each noncertainty stratum, a simple random sample of EINs was selected. The sampling rates for the EINs selected in the sample for the annual survey varied between 1 in 1 and 1 in 400.

Sample Maintenance

Periodically, we update the sample to represent EINs issued since the last sample selection. For all EIN "births," a two-phase selection procedure is used. EIN births are new EINs recently assigned by the IRS, on the latest available IRS mailing list for FICA taxpayers, and assigned a kind-of-business classification (if possible) by the Social Security Administration (SSA). In the first phase, births are stratified by kind of business and size (expected employment or quarterly payroll). A relatively large sample is drawn and canvassed to obtain a more reliable measure of size, consisting of sales in two recent months, and a more detailed kind-of-business code, if needed.

Using this more reliable information, the selected births from the first phase are subjected to probability-proportional-to-size sampling with overall probabilities equivalent to those used in drawing the initial sample from the 1999 Business Register. Because of the lag in reporting births to the IRS, and the time needed to accomplish the two-phase birth-selection procedure, births are added to the sample about nine months or more after they begin operation.

Births that are selected in the quarterly birth-selection procedures through November of the annual survey year are included in the main mailing of the ATS questionnaires in January of the following year. To better represent all EIN births in the annual survey year, and specifically to account for the coverage lag in the birth-selection procedure, we add births to the annual sample that are selected in February, May, and August of the year following the survey year. We mail survey forms to these births in June and August to supplement the main annual survey mailing.

To be eligible for the sample canvass and tabulation in a given month, a wholesale EIN selected in the noncertainty sampling operations must meet both of the following requirements:

! It must be on the latest available IRS mailing list for FICA taxpayers from the previous quarter.

! It must have been selected from either the Business Register or the file of employer births.

EINs selected into the sample with certainty are not dropped from canvass and tabulation if they are no longer on the IRS mailing list. Rather, the business that used the EIN is contacted, and if a successor EIN is found, it is added to the survey. This is a more stringent quality control used for these larger businesses.

Annual Estimation Procedures

The estimates published in this report are derived from the Annual Trade Survey. The estimates are computed as the sum of weighted data (reported and imputed) for all selected sampling units that meet the tabulation criteria given in the **Sample Maintenance** section. The weight for a given sampling unit is the inverse of its probability of selection.

Census Adjustment Procedures

The estimates in this report have been linked to the 1997 Economic Census to reduce sampling error and to allow comparability with the census using the following procedure. Unadjusted estimates of total sales and e-commerce sales were formed by summing weighted sampling unit data (both reported and imputed). Estimated total sales for 1998 for detailed NAICS kinds of businesses were then adjusted to the 1997 Economic Census sales totals. For each detailed kind of business, the ratio of the adjusted 1998 sales to the unadjusted 1998 sales was multiplied by the unadjusted e-commerce sales to derive an adjusted e-commerce estimate. The detailed e-commerce estimates were summed to produce the estimates in this report.

Dollar Values

All dollar values presented are expressed in current dollars; that is, the estimates are not adjusted to a constant dollar series. Consequently, when comparing estimates to prior years, users also should consider price level changes.

Census Disclosure Rules

In accordance with Federal law governing Census Bureau reports, no estimates are published that would disclose the operations of an individual firm.

RELIABILITY OF THE ESTIMATES

An estimate based on a sample survey potentially contains two types of errors - sampling and nonsampling. Sampling error occurs because characteristics differ among sampling units and because only a subset of the entire population is measured in a sample survey. Nonsampling error encompasses all other factors that contribute to the total error of a sample survey estimate. The accuracy of a survey result may be affected by these two types of errors.

Sampling and nonsampling errors are often measured by the quantities, bias and variance. The *bias* of an estimator of an unknown population value is the difference, averaged over all possible samples of the same size and design, between the estimator and the unknown population value. Any systematic error, or inaccuracy that affects all samples of a specified design in a similar way, may bias the resulting estimates. The *variance* of an estimator is the squared difference, averaged over all possible samples of the same size and design, between the estimator and its average value.

Descriptions of sampling and nonsampling errors for the ATS are provided in the following sections.

Sampling Error

Because the estimates are based on a sample, exact agreement with results that would be obtained from a complete enumeration of firms represented on the sampling frame using the same enumeration procedures is not expected. However, because each firm on the sampling frame has a known probability of being selected into the sample, it is possible to estimate the sampling variability of the survey estimates.

The particular sample used in this survey is one of a large number of samples of the same size that could have been selected using the same design. If all possible samples had been surveyed, under the same conditions, an estimate of an unknown population value could have been obtained from each sample. These samples give rise to a distribution of estimates for the unknown population value. A statistical measure of the variability among these estimates is the standard error, which can be approximated from any one sample. The *standard error* is defined as the square root of the variance. The *coefficient of variation* (or relative standard error) of an estimator is the standard error of the estimator divided by the estimator. Note that measures of sampling variability, such as the standard error or coefficient of variation, are estimated from the sample and are also subject to sampling variability. (Technically, we should refer to the *estimated* standard error or the *estimated* coefficient of variation of an *estimator*. However, for the sake of brevity we have omitted this detail.) It is important to note that the standard error and coefficient of variation only measure sampling variability. They do not measure any systematic biases in the estimates. Estimated coefficients of variation for dollar volume estimates and estimated ratios are shown in Table 2A. (All coefficients of variation are expressed as percents.)

The estimate from a particular sample and the approximate standard error associated with the estimate can be used to construct a confidence interval. A *confidence interval* is a range about a given estimator that has a specified probability of containing the estimator's corresponding, unknown population value. Associated with each interval is a percentage of confidence, which is interpreted as follows. If, for each possible sample, an estimate of an unknown population value and its approximate standard error were obtained, then:

1. For approximately 90 percent of the possible samples, the interval from 1.645 standard errors below to 1.645 standard errors above the estimate would include the unknown population value.

2. For approximately 95 percent of the possible samples, the interval from two standard errors below to two standard errors above the estimate would include the unknown population value.

To illustrate the computation of a confidence interval for an estimate of total revenue, assume that an estimate of total revenue is $10,750 million and the coefficient of variation for this estimate is 1.8 percent, or 0.018. First obtain the standard error of the estimate by multiplying the total revenue estimate by the coefficient of variation. For this example, multiply $10,750 million by 0.018. This yields a standard error of $193.5 million. The upper and lower bounds of the 90-percent confidence interval are computed as $10,750 million plus or minus 1.645 times $193.5 million. Consequently, the 90-percent confidence interval is $10,432 million to $11,068 million. If corresponding confidence intervals were constructed for all possible samples of the same size and design, approximately 9 out of 10 (90 percent) of these intervals would contain the unknown population value.

Nonsampling Error

Nonsampling error encompasses all other factors that contribute to the total error of a sample survey estimate and may also occur in censuses. It is often helpful to think of nonsampling error as arising from deficiencies or mistakes at some point in the survey process. In the ATS, nonsampling error can be attributed to many sources: inability to obtain information about all units in the sample; response errors; differences in the interpretation of the questions; mistakes in coding or keying the data obtained; and other errors of collection, response, coverage, and processing. Although no direct measurement of the potential biases due to nonsampling error has been obtained, precautionary steps were taken in all phases of the collection, processing, and tabulation of the data in an effort to minimize their influence.

A potential source of bias in the estimates is nonresponse. Nonresponse is defined as the inability to obtain all the intended measurements or responses about all selected firms. Two types of nonresponse are often distinguished. *Unit nonresponse* is used to describe the inability to obtain any of the substantive measurements about a sampled firm. In most cases of unit nonresponse, the questionnaire was never returned to the Census Bureau, after several attempts to elicit a response. *Item nonresponse* occurs either when a question is unanswered or the response to the question fails computer or analyst edits.

For both unit and item nonresponse, a missing value is replaced by a predicted value obtained from an appropriate model for nonresponse. This procedure is called *imputation*. Imputed sales amounts to about 7.7 percent of the total sales estimate. Imputed e-commerce sales amounts to about 9.5 percent of the total e-commerce sales estimate.